# From Data Communication to Delivery of Quality Data

Leon Reznik, Igor Khokhlov
Department of Computer Science,
Rochester Institute of Technology
E-mail: lr@cs.rit.edu

## 1. Big Data revolution requires integration of data management and network services.

The advances in the information and communication technologies over the last decade laid a strong foundation for data generation and storage on a staggering scale. Also, this development moved the point of possible data use far away from data source locations, thus enhancing the importance of cybersecurity and safety factors consideration in system design. Progress in nanoscale devices, Internet of Things, Citizen Science made possible to collect huge volumes of the data of wide varieties generated with high velocity. This data may have rather poor quality characteristics, such as low signal-to-noise ratio, high probability of errors, high distortion, dynamic device non-uniformity. If the current trend continues, we may expect an unprecedented scale of generating, storing, and communicating more and more data of low quality overfilling the available capacities. The current data management structures would not be able to achieve good data value with high levels of confidence, trustworthiness, accuracy, reliability, security, and safety. Without significant changes, existing infrastructure will not scale up these features to huge data arrays, which will have to be communicated, computed, and controlled. *Novel data management principles that should involve processing and filtering the data based on their quality characteristics need to be developed and employed. They have to be supported by network services and protocols facilitating delivery not data only but also data quality (DQ) characteristics (see fig.1).*

Due to many concern, *the US National Academy of Sciences* and other professional and government organizations recently introduced new initiatives aiming at attracting more attention and research efforts to DQ issues, which are critical for ensuring the soundness of the scientific endeavor itself, given the crucial role that DQ plays in the ability of scientists to reproduce research findings, as reiterated by the US National Academy of Sciences Committee on Ensuring the Utility and Integrity of Research Data in the Digital Age.

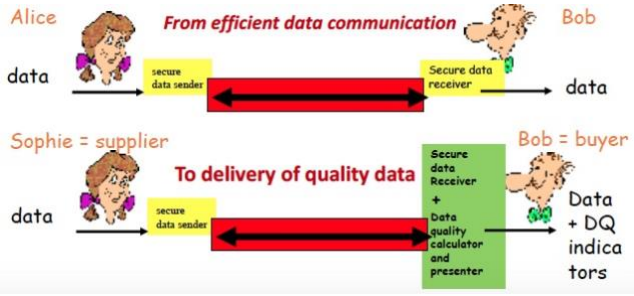


Figure 1. Changes in Network Services model to be discussed

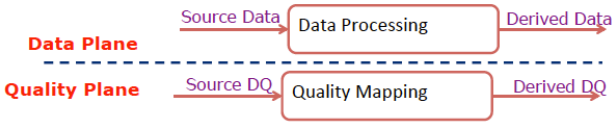


Figure 2. Novel multidimensional data processing and communication

## 2. Goal of this research and development

**Our major goal is to develop methodologies and technologies that will provide an end user or an application with data of a specified quality at the point of use.** This goal could be achieved by the dynamic selection, preferably in real-time, of the data sources and communication paths from them to the points of data use. Sources might include physical sensors, data generators and crowd-sources, storage facilities, cloud services, etc. DQ indicators are composed from multiple diverse metrics integrating a wide variety of characteristics such as accuracy, reliability, timeliness, security, and safety. This innovation has a high potential to enable a significant improvement in a wide spectrum of science and technology applications as it will create new opportunities for optimizing data structures, data processing and fusion procedures based on a new quality and security information usage. By presenting to an end-user or an application the integral DQ indicators, which include cybersecurity evaluation, it will significantly increase the system's trustworthiness and confidence, with which users interact with infrastructure. Advancing the design of high-confidence systems will transform the ways of interacting with an engineered system by allowing a user to understand and compare various data files, streams, and sources based on the associated DQ with integral quality characteristics reflecting various aspects of a system's functionality. It will create a foundation to transform one-dimensional (data only) data collection and processing structures into multi-dimensional (data plus DQ indicators) procedures (fig. 2). This approach will allow moving *from extensive infrastructure development* that involves generating huge amounts of data that need to be communicated, stored, and processed *to intensive services development* that will generate and deliver data when and where there is a need for it.

## 3. Goal from the network perspectives.

*The proposed approach will integrate data management and network services.* DQ at the point of use will depend on multiple factors, including the DQ at the source and quality of network services among the most influencing ones. Quality of service (QoS) demonstrates the overall network performance, particularly the performance seen by the users of the network. To quantitatively measure the quality of service, several related aspects of the network service are

often considered, such as error rates, bit rate, throughput, transmission delay, availability, jitter, etc. While various Internet models have been discussed, some of them (e.g., Internet2 group) chose not to deploy QoS protocols inside its Abilene Network and instead to aim at providing over-positioning of capacity and bandwidth. The approach proposed in this paper does not include prioritizing services and/or applications by providing better conditions and more resources. It does not contradict to the Internet2 model but rather builds upon it as well as Internet Engineering Task Force (IETF) end-to-end QoS developments such as Next Steps in Signaling (NSIS) and End-to-end Quality of Service over Heterogeneous Networks (EuQoS) and other European projects.

## 4. Proposed solutions: DQ evaluation and assurance framework

To achieve our goal formulated in sec. 2, we have to solve two interrelated problems:

1. To develop the techniques and tools for DQ evaluation.
2. To develop the techniques and tools for DQ assurance at the point of data use.

In our NSF funded research (award ACI-1547301), we are building a proof-of-the-concept design, which will be used to develop, verify and promote a comprehensive methodology for DQ evaluation focusing on an integration of cybersecurity with other diverse metrics reflecting DQ, such as accuracy, reliability, timeliness, and security into a single methodological and technological framework.

The brief framework operation is presented in fig. 3. Steps 1-3 describe the process of deriving data quality, ranging from the original data collection, and transformation of data using parsers into a standard format, which is appropriate and suits the framework, calculating and integrating the DQ metrics and finally displaying results based on the dataset. Steps 4-5 aim at ensuring the required DQ at the point of data use by dynamic reselection of data sources and data delivery routes. The framework design is presented in our webinar recordings available online [1]. For the framework implementation on Android mobile devices, a number of applications for evaluating data quality and security have been produced and made available on Google Play Store. The framework was validated on a number of test cases [2-5]. The machine learning based procedures for the automatic sensor choice and its implementation with the mobile device sensors have been developed too [6].

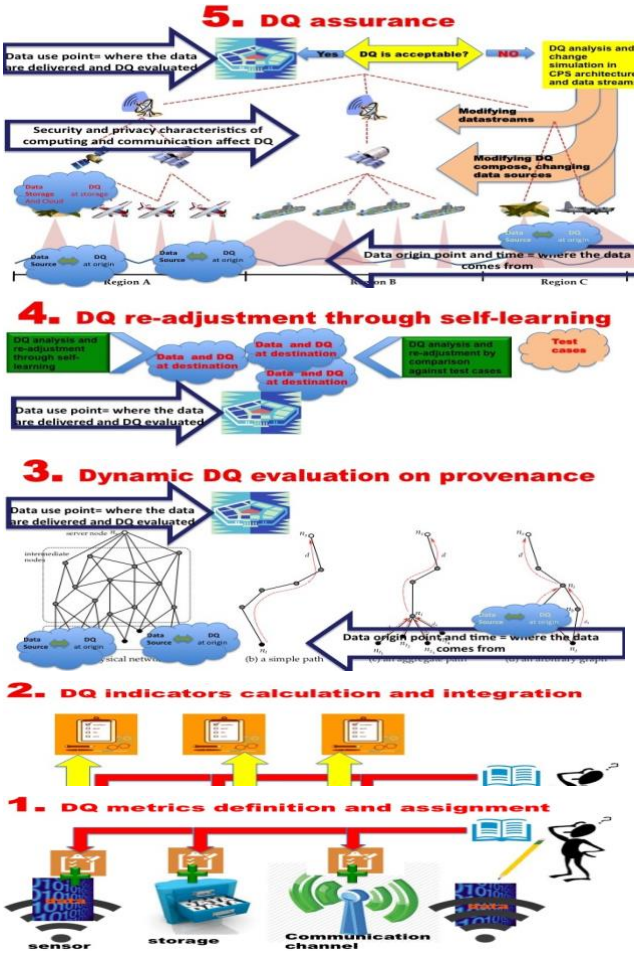


Figure 3. DQ evaluation and assurance framework operation: 1. DQ metrics composition and assignment, 2. DQ initial indicators calculation based on the metrics and their integration, 3. Dynamic DQ changes based on data provenance, 4. DQ integral indicators re-adjustment and 5. DQ assurance procedures

## References


[1] L. Reznik and I. Khokhlov "Data Quality & Security Evaluation Framework Development" presented at the National Center for Supercomputing Applications under auspices of NSF, on March 26, 2018 , recordings available on the YouTube channel https://www.youtube.com/watch?v=nkp0kvJvTWw&feature=youtu.be.

[2] I.Khokhlov and L. Reznik, "Android system security evaluation," 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2018, pp. 1-2.

[3] A.Vora, L. Reznik and I. A. Vora,L.Reznik, I.Khokhlov, "Mobile road pothole classification and reporting with data quality estimates," 2018 Fourth International Conference on Mobile and Secure Services (MobiSecServ), Miami Beach, FL, 2018, pp. 1-6.

[4] I.Khokhlov, L. Reznik, A. Kumar, A. Mookherjee, R. Dalvi Data Security and Quality Evaluation Framework: Implementation Empirical Study on Android Devices, 2017 20th IEEE Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT), St.Petersburg, April 3-7, 2017, pp. 161-168

[5] I.Khokhlov, L. Reznik Data Security Evaluation for Mobile Android Devices, 2017 20th IEEE Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT), St.Petersburg, April 3-7, 2017, pp. 154-160

[6] I.Khokhlov, A. Pudage and L. Reznik, "Sensor selection optimization with genetic algorithms," *Proc. IEEE Conf. on Sensors*, Montreal, Canada, October 27-30, 2019.